The Role of Replication in Psychological Science

Samuel C. Fletcher*

Department of Philosophy
University of Minnesota, Twin Cities

August 26, 2020

Abstract

The replication or reproducibility crisis in psychological science has renewed attention to philosophical aspects of its methodology. I provide herein a new, functional account of the role of replication in a scientific discipline: to undercut the underdetermination of scientific hypotheses from data, typically by hypotheses that connect data with phenomena. These include hypotheses that concern sampling error, experimental control, and operationalization. How a scientific hypothesis could be underdetermined in one of these ways depends on a scientific discipline's epistemic goals, theoretical development, material constraints, institutional context, and their interconnections. I illustrate how these apply to the case of psychological science. I then contrast this "bottom-up" account with "top-down" accounts, which assume that the role of replication in a particular science, such as psychology, must follow from a uniform role that it plays in science generally. Aside from avoiding unaddressed problems with top-down accounts, my bottom-up account also better explains the variability of importance of replication of various types across different scientific disciplines.

1 Introduction: The Replication Crisis

According to a recent survey of 1,576 scientists conducted by the preeminent general science journal *Nature*, there is strong evidence that the majority of researchers believe that there

*Thanks to audiences in London (UK XPhi 2018), Burlington (Social Science Roundtable 2019), and Geneva (EPSA2019) for their comments on an earlier version, and especially to the Pitt Center for Philosophy of Science Reading Group in Spring 2020: Jean Baccelli, Andrew Buskell, Christian Feldbacher-Escamilla, Marie Gueguen, Paola Hernandez-Chavez, Edouard Machery, Adina Roskies, and Sander Verhaegh.

is a replication or reproducibility crisis in science. Indeed, "More than 70% of researchers have tried and failed to reproduce another scientist's experiments, and more than half have failed to reproduce their own experiments" (Baker, 2016). The survey results suggest a widespread malaise, albeit with some variability between disciplines and concentration in the biomedical and social sciences—e.g., in cancer research (Begley and Ellis, 2012; Lawrence et al., 2013; Nosek and Errington, 2017), neuroscience (Button et al., 2013), and (to a lesser extent) experimental economics (Camerer et al., 2016). But these concerns have, beyond any other scientific discipline, focused on psychology, which is the focus of the present essay. For example, in a study that has since received widespread attention, the Open Science Collaboration (OSC) had just before (2015) attempted to replicate 100 social and cognitive psychology experiments, coming to the "same result" only one-third to one-half of the time, depending on the criterion for replication used.

Interpreting the significance and sources of these results is complex and ongoing, but they and other related events have certainly renewed attention to philosophical aspects of methodology in psychological science.[1] Most psychologists writing on the subject simply affirm the importance of replication in science simpliciter, drawing the implication that if psychological science is to live up to its name, it must engage in and take seriously the results of various replication studies. For example, Simons (2014, p. 76) affirms that "Reproducibility is the cornerstone of science" and Rosenthal (1990, p. 2) that "Scientists of all disciplines have long been aware of the importance of replication to their enterprise." Sometimes, their affirmations come with appeals to the authority of methodologists and philosophers of science. "Reproducibility is a defining feature of science" and "a core principle of scientific progress" write the Open Science Collaboration (OSC) (2015, p. 1), citing (but not quoting) psychologists Meehl and Platt and philosophers as varied as Hempel, Lakatos, and Salmon. Scientific progress here refers the process of accumulation of evidence for scientific claims and the establishment of scientific discoveries. Citing Popper (1959, pp. 23–24),[2] Zwaan et al. (2018, p. 2) write that "There are two important aspects to these insights [of Popper's]

[1]These related events include Daryl Bem's use of techniques standard in psychology to show evidence for extra-sensory perception (2011), the revelations of high-profile scientific fraud by Diederik Stapel (Callaway, 2011) and Marc Hauser (Carpenter, 2012), and related replication failures involving prominent effects such as ego depletion (Hagger et al., 2016).

[2]The quotation reads: "the scientifically significant physical effect may be defined as that which can be regularly reproduced by anyone who carries out the appropriate experiment in the way prescribed." See also Popper (1959, p. 45): "Only when certain events recur in accordance with rules or regularities, as in the case of repeatable experiments, can our observations be tested—in principle—by anyone. ...Only by such repetition can we convince ourselves that we are not dealing with a mere isolated 'coincidence,' but with events which, on account of their regularity and reproducibility, are in principle inter-subjectively testable." Zwaan et al. (2018, pp. 1, 2, 4) also quote Dunlap (1926) (published earlier as Dunlap (1925)) for the same point.

that inform scientific thinking. First, a finding needs to be repeatable to count as a scientific discovery. Second, research needs to be reported in such a manner that others can reproduce the procedures" used to generate the finding that the research reports.

I am sympathetic to their desire to provide a rationale for the importance of replication and a justification for its role in psychological science. To do so, however, instead of beginning with assumptions about general scientific virtues such as epistemic progress or objectivity, I focus on specific underdetermination problems that empirical investigations face. These problems concern the underdetermination of scientific hypotheses from data, typically not by other scientific hypotheses, but by hypotheses that connect data with phenomena. Such hypotheses assert, for example, the data's representativeness of the population from which it is drawn, or the internal validity of the constructed psychological assessment. Data not being representative or a failure of internal validity can often provide explanations of the results of scientific studies alternative to the substantive hypotheses the studies seem to support. I describe these in more detail in section 2, drawing on the functional classification of replication efforts by Schmidt (2009, 2017).[3]

Besides setting the terms of the debate more precisely, this also makes explicit the variety of types of underdetermination that replication can address in scientific inquiry—psychology, in particular. I argue in section 3 that how central replication is to a given scientific discipline depends on a confluence of epistemic, material, and institutional factors particular to it, such as its goals, the availability and cost of obtaining evidence, and the organization structure of its research community. In the case of psychological science, importantly, these factors' *interactions* contribute to replication's essential role therein to ameliorate the underdetermination challenges it constantly faces.

This is a "bottom-up" account of the role of replication in psychological science because it begins with delineating the particular, often low-level underdetermination challenges which that science faces. (As I describe in more detail in section 2, "low-level" refers to data, phenomena, or experiment rather than theory.) For contrast, I examine in section 4 one of the further developed "top-down" accounts—those that begin with principles about (good) science in general—more popular among psychologists, by Zwaan et al. (2018). They draw on Lakatos' Methodology of Scientific Research Programmes (MSRP) (1970) and his "sophisticated falsificationism." According to MSRP, one can identify good science—a *progressive*

[3]Schmidt (2009, pp. 90–2), citing much the same passages of Popper (1959, p. 45) as the others mentioned, also provides a similar explanation of replication's importance, appealing to general virtues such as objectivity and reliability. (See the first paragraphs of Schmidt (2009, p. 90) and Schmidt (2017, p. 236) for especially clear statements, and Machery (2020) for an account of replication based on its ability to buttress reliability in particular.) But for him, that explanation only motivates why establishing a definition of replication is important in the first place; it plays no role in his definition itself. Thus, by drawing on Schmidt's account of *what* replication is, I am not committing to his and others' stated explanations of *why* is important.

research programme—according to the extent that new theoretical content is added cumulatively over time to a fixed "hard core" of theory, which also adds cumulatively corroborated predictions. Conversely, science is bad—a *degenerating* research programme—to the extent that it instead encounters repeated falsification of such new theoretical content and corroborates few new predictions. Zwaan et al. (2018, p. 2) suggest that "replications are an instrument for distinguishing progressive from degenerative [i.e., degenerating] research programs." Progressive research programs have successfully replicated experiments; degenerating ones have failed replicated experiments. Replication provides a means to distinguish which parts of psychology are good and progressive from those that are bad and degenerating.

However, MSRP doesn't support the employment of replication or its importance in science generally or even in psychology specifically. There are at least three reasons for this. Two of them involve longstanding problems for MSRP. The first is to describe adequately scientific theories that use probability in their theorizing or models essentially, as is common in most of quantitative psychology. The second is to have clear guidance on MSRP's normative consequences—how in practice to determine which research programs are progressive and which are degenerating. The third observes that the account of replication's importance using just MSRP seems to demand that the function and role of replication across all good science is the same. Yet, it is not difficult to find sciences where the types of replication (if any) considered to be relevant are quite different from those in psychology.

In the final section 5, I draw three concluding comparisons between bottom-up and top-down approaches. The first elaborates on the just-mentioned point: bottom-up approaches seem to account better for the diverse importance and roles of various types of replication than top-down ones. Second, bottom-up approaches do not entail anything about how universal replication's importance is. Third, both approaches afford unity to understanding the role of replication but at different levels: top-down approaches derive replication's role in a particular science from its role in science generally, while my bottom-up account unifies them functionally, even if that function is realized in different ways in different scientific investigations.

2 The Functions of Replication

There are may typologies for replication, understood broadly as the repetition of some scientific research procedure (Gómez et al., 2010; Fidler and Wilcox, 2018, §1). Some of them use the terms "replicability," "reproducibility," "repeatability," etc., interchangeably, and others use them to denote distinct types. In this essay, I will attempt to use "replicability" and related declensions exclusively. (Any references to "reproducibility," for example, should be

understood as exchangeable with "replicability.")

I will also adopt as my primary classification the functions for replications as elaborated by Schmidt (2009, 2017). Each function is to ameliorate a particular type of underdetermination of a tested scientific hypothesis, sometimes by other (slightly different) scientific hypotheses but mostly by low-level hypotheses about, e.g., statistical analysis and novel operationalizations of theoretical concepts pertaining to phenomena or experiment. Thus, by "low-level," I mean "pertaining to (models of) data or phenomena or experiment" in contrast with ("high-level") theory, invoking the data-experiment-theory hierarchy of Suppes (1962, 2007) or the data-phenomena-theory hierarchy of Bogen and Woodward (1988). Theory makes predictions about experiments or phenomena, while data are the closest manifestations of our observations or measurements. One must assume hypotheses about each level in the hierarchy and how these levels interact in order to connect observation with scientific theory. Which one of these hierarchies is preferred and the details of how they should be elaborated is not so relevant for present purposes because these matters do not play a significant role in my account of the function of replication.[4] These hierarchies and the "low-level" appellation rather mainly serve to distinguish the sort of underdetermination commonly at issue here with a perhaps more familiar ("high-level") sort pertaining to scientific theory and its implications for scientific realism (Stanford, 2017).[5]

How, then, does each function of replication serve to ameliorate underdetermination? It does so simply by disconfirming certain competing hypotheses, or alternately, confirming their negations as auxiliary hypotheses used in the interpretation of scientific results. Each function is therefore to confirm the one of the following six claims about the original results:[6]

1. *They are not due to mistakes in the data analysis.* This is a low-level hypothesis about the data. Ensuring that the results of a scientific study are not due to data-entry, programming, or other suchlike technical errors often involves re-calculating the values

[4]For example, it is compatible with modifications or clarifications of how interpretation plays an essential role in determining what data models are or what they represent, either for Suppes' hierarchy (Leonelli, 2019) or Bogan and Woodward's (Harris, 2003). It is also compatible with interactions between the levels of data and phenomena (or experiment) in the course of a scientific investigation (Bailer-Jones, 2009, Ch. 7).

[5]That's not to say there is *no* interesting relationship between low-level underdetermination and the question of scientific realism, only that it much more indirect. See Laymon (1982) for a discussion thereof and Brewer and Chinn (1994) for historical examples from psychology as they bear on the motivation for theory change.

[6]The first function, concerning mistakes in data analysis, does not appear in Schmidt (2009, 2017). That said, neither he nor I claim that our lists are exhaustive, but they do seem to enumerate the most common types of low-level underdetermination that arise in the interpretation of the results of psychological studies. One type that occurs more often in the physical sciences concerns the accuracy, precision, and systematic error of an experiment or measurement technique; I hope in future work to address this other function in more detail. It would also be interesting to compare the present perspective to that of Feest (2019), who, focusing on the "epistemic uncertainty" regarding the third and sixth functions, arrives at a more pessimistic and limiting conclusion about the role of replication in psychological science.

of statistics, ensuring the validity of computer programs and the implementation and consistency of data processing algorithms, etc. This may include checking that the results remain the same when alternative analyses are preformed, if there were any conventional aspects in the original analysis (Nuijten et al., 2018).

2. *They are not due to sampling error.* This is low-level a hypothesis about the experiment. Essentially every study using statistical models has a probability of producing data that constitute misleading evidence for a hypothesis or theory. Typically, the more data collected, the lower the probability of this occurring.

3. *They do not depend on contextual factors, according to the theory or hypothesis tested.* This is a hypothesis about the phenomena and theory. Sometimes not all contextual variables or interactions between those variables and the independent variables of a study are accounted for in the data analysis of a study, while accounting for them would change the study's support for the scientific hypothesis of interest. What these variables or interactions could be depends on the theory or hypothesis tested. This function is sometimes regarded as for the *internal validity* of the techniques used in the original study.

4. *They do not arise from fraud or questionable research practices.* This is a low-level hypothesis about the data and experiment. The soundness of a study's results depends on it accurately representing the methods used. Questionable research practices (and, more extremely, fraud) undermine this soundness by misrepresenting the warrant for the methods presented. For example, conclusions based on statistical inferences from data may not be warranted if the data was fabricated or the inferences presented do not match the inferences undertaken.

5. *They generalize, according to the theory or hypothesis tested, to a larger or different population than that sampled in the original.* This is a hypothesis about the phenomena and theory. Oftentimes the theory or hypothesis that a study tests pertains to a larger population than that from which the study actually drew. Thus its results only bear on the tested subpopulation without further assumptions.

6. *Their aspects pertaining to the theoretical hypothesis of interest hold even when that hypothesis is operationalized or tested in completely different ways.* This is a hypothesis about the phenomena. To fulfill this function, a replication must explore implications of the hypothesis tested that go beyond that of the original study, even if all previously listed claims have been supported.

For a study to fulfill a certain replication function best, it makes only certain variations on the original study that it replicates. Cross-cutting these variations is who the *replicator* is—the person or group of people undertaking a replication (Radder, 1992). It could be the same as or different from that for the original study, one of which is sometimes more apt for a given function than the other.

The first function, ruling out mistakes in the data analysis, is the strictest in a sense because it demands that the replication study be performed on the same data set as the original study. Hence, it is often called *methods* replication. Typically, with an eye towards the fourth function of ruling out fraud and questionable research practices, a research team different from the one that performed the original study should perform a methods replication. Methods replication is important in any scientific study with non-trivial data analysis, but there seems to be little controversy about this (Zwaan et al., 2018, p. 48).[7]

Setting the function of methods replication aside, the remaining five functions involve two experiments or scientific studies, an original and an attempted replication, each with distinct or largely distinct data or observations. The second, which is the replication, varies at least one of the following four classes of variables with respect to the first (Schmidt, 2009, p. 93):

1. the procedures for constituting the independent variables, the ones whose explanatory, predictive, or controlling features are under test;

2. the study's context, i.e., possible moderators such as the properties and history of the research units and the people running the study, their relevant historical or cultural context, and the physical setting of the study and its material realization;

3. the procedures for the selection and allocation of the research units; and

4. the procedures for constituting the dependent variables, the ones to be explained, predicted, or controlled.

The minimal variation on an original study in a replication to gain evidence against its result being due to sampling error is to keep all of these variable the same, with the exception of selecting different token research units (albeit using the same selection procedures). The minimal variation against its result being due to uncontrolled variation of contextual factors is only to change the context in an explicit way. (Both of these variations can test against fraud but demand a different group running the study than the original.) The minimal variation to show that a result generalizes according to the theory tested is to expand or

[7]For examples from economics, see Cartwright (1991, pp. 145–6); for examples from gravitational and particle physics, see Franklin and Howson (1984, pp. 56–8).

change the selection of participants. But to show that the result generalizes regardless of operationalization, variations on most or all of the above variables may be necessary.

Thus replications might be grouped into two categories depending on the minimal variations needed to perform certain replication functions. Schmidt (2009, p. 91) calls those with functions 2–5 direct replications, and those with function 6 conceptual replications, noting that, even though the terminology is his, psychologists have proposed similar typologies since at least the 1960s. Philosophers have for some time also noted a functional difference between (something like) these types of replication (Franklin and Howson, 1984; Cartwright, 1991; Radder, 1992). That said, replication studies that do not follow the minimum variations prescribed above may be classified as neither direct nor conceptual. Direct and conceptual replications lie on a continuum of similarity with respect to the original experiment, with direct replications more similar and conceptual replications less similar (LeBel et al., 2017; Rosenthal, 1990). For these and other reasons, the distinction between direct and conceptual replication has been more controversial lately (Nosek and Errington, 2020; Machery, 2020). But for present purposes, that distinction is just a convenient overlay of some "landmarks" within the replication topography, so I use these labels in the sequel just to denote typical cases. What is more fundamental is the underlying classification of replication efforts by their functions, namely to ameliorate specific underdetermination problems regarding the interpretation of an original study's results.

One might object, however, that there is too much flexibility of interpretation to locate a particular experiment on this continuum objectively, hence objectively identify a replication's function. For example, Gelman (2018, p. 19) writes that due to the inevitable numerous circumstantial differences between any original study and a replication, the latter "could be considered a 'direct replication' if that interpretation is desired, or a mere 'extension' or 'conceptual replication' if the results do not come out as planned." The idea is that researchers could interpret any circumstantial differences between studies as part of the contextual set that makes a difference to the predictions drawn from the hypothesis or theory under study. For example, should "college students" be considered the population from which a study conducted at a Bavarian university sampled? Should it be more specific, a distinct population from one conducted in Berlin? And what of all the enumerable contextual factors inevitably at play in social research, such as the time of day and year or the idiosyncrasies of the rooms where data was taken?

This sort of concern, I insist, is based on an understandable misinterpretation of the relata of the replication concept and the appropriate constraints on hypothesis specification. As Franklin and Howson (1984, p. 53) pointed out, judgments about the similarity of two experiments is always *relative* to one or more theories of those experiments' functioning.

Direct replications use much the same such theory with the same set of variables, while conceptual replications use largely distinct such theories. This is despite the linguistic fact that "is a replication of" suggests that the concept is a binary relation between experiments. So regarding a replication as direct or conceptual—or more generally, as being towards one function or another—is not merely a change in "interpretation" or something that can be chosen conventionally or at will, but a change in the very theory under test. Thus the results of any replication provide evidence about the constraints on the scope and generality of psychological theories and hypotheses, i.e., what populations they pertain to and which contextual factors they are independent of.

Now, it's true that published account of studies do not—indeed, cannot—include *all* of these constraints and possible factors. But this needn't pose a problem either if one adopts the reasonable assumption, long propounded within the psychological research literature, that unless some factor or condition is specified in a study, researchers are, ceteris paribus, justified in assuming that the study hypothesizes that the results are independent of an such unlisted contextual factor or condition (Greenwald et al., 1986; Simons et al., 2017). Thus any further studies that measure the effects of previously uncontrolled contextual variables explore the empirical tenability of that independence.

3 A Bottom-Up Account of the Role of Replication

I aver that replication is important in psychological science because of its role in addressing the often low-level sorts of underdetermination problems reviewed in the previous section 2. In order to support this conclusion, however, I must show how at least some of these problems actually arise in psychological studies. It is not apparently an a priori matter, nor do these problems arise equally in all sciences or in the same way. But, "if replication is not a universal ideal, then where do we draw the line? How can we know to which fields the norm applies and to which it doesn't?" (Guttinger, 2020, p. 9). This question arises for Guttinger in the context of describing and endorsing what he calls "the new localism in the replication crisis debate [which] targets the idea of replicability as a universal standard" (2020, p. 6).

I, however, decline drawing a line at all—that is, giving necessary and sufficient conditions for which fields demand replication and which don't. One reason is that this demarcation problem seems to be as intractable as the traditional one of separating science from non-science or pseudo-science (Laudan, 1983). This difficulty notwithstanding, another is that it presupposes any line must be divided by field: if replications' functions are to ameliorate underdetermination problems, then the relevant unit of analysis over which replication is important or not is much finer, on the level of scientific hypotheses. A third difference,

which I discuss further in section 5, is that I remain agnostic regarding whether replication of some form is important for all scientific endeavors (i.e., a "universal standard" of sorts).

Nevertheless, I agree with part of the spirit of his localism, namely that we might better understand where and why replication has an important role by identifying "aspects of scientific practice that can [bear upon replication's role]: the type of questions addressed, the setup used, and the nature of the objects analyzed" (Guttinger, 2020, p. 14). Consequently, I will instead proceed to describe a collection of conditions that are jointly sufficient to justify why certain types of underdetermination arise routinely in psychology, hence why replication is important in psychology generally (if not in each and every of its investigations). Each of them concerns epistemological, material, or institutional features particular to contemporary psychological science:[8]

1. The kinds of questions asked and the nature of the subject matter.
2. The experimental and data analytic techniques used.
3. The nature, complexity, and diversity of the objects studied.
4. The state and development of overriding theory to constrain possibilities.
5. The cost and availability of evidence.
6. The institutional features of academic psychology, especially the distribution of resources and structure of incentives.

These conditions (or analogs thereof) likely bear on the types of underdetermination, hence the importance of replication, found in other scientific disciplines. Further conditions may also bear, delineation of which I leave to future investigation.

Although there are connections between each of these conditions that bolster the importance of replication in psychology, there are some connections that are stronger than others. In what follows I will therefore present them roughly in three pairs in the order presented above—i.e., 1 and 2, 3 and 4, and 5 and 6—focusing on the functions for replication (discussed in §2) that they entreat, beginning with the kinds of questions asked, the nature of the subject matter, and the experimental and data analytic techniques used.

Psychology *is*—or, at least, *aspires to be*—a science that seeks to describe, explain, and predict stable patterns of human behavior and mental life rather than singular events. Accordingly, its experiments on and observations of these stable patterns ought to deliver similar results. Cartwright (1991, pp. 146–7) put the point well about physics and economics:

[8]This is also analogous to the case of the demarcation problem, on which progress might be possible if one helps oneself to discipline-specific information (Hansson, 2013).

> We are not just trying to measure some very local occurrence—a fact about the electrons in this iron bar or about the savings patterns in just the individuals sampled over just the period considered. Rather we are trying to measure some general phenomenon. ...We expect therefore that the instrument must produce the same results wherever it can be applied to look at the very same phenomenon. It should, as it were, give the same results on different samples. When it doesn't, either your hypothesis or the instrument must be faulted.

Thus, to test hypotheses in psychological science, one must be able to test one's instruments, which is a question of the internal validity of the experimental procedure and one of the functions of direct replication. These procedures measure quantitative information such as the strength or frequency of responses from individuals according to some prescribed stimulus. In particular, insofar as the quantitative nature of this data facilitates analysis via statistical models, it suggests the need for direct replication addressing questions of sampling error.[9]

There are many reasons for the statistical nature of data models in psychology, even for many qualitative forms of data (Suppes, 2007). Two that are especially relevant here concern the nature, complexity, and diversity of human subjects, and the relatively underdeveloped state of psychological theory. (These are the third and fourth conditions listed above.) Moreover, these interact. The complexity and diversity of human subjects refers to the great variety and interconnectedness of factors that could be relevant in affecting human behavior and mental life. This means that there are potentially very many unexpected ways that true effects and merely apparent ones could arise, and that in many cases, measured responses result from the accumulated effects of many small unknown causes, a situation paradigmatic for statistical modeling. This motivates the need for evidence about the internal and external validity of a research hypothesis, hence certain direct and conceptual replications, including those about the generality of the human population for which the hypothesis holds and about experimental sampling error.

In principle, strong, unified psychological theory would provide both precise predictive hypotheses and a lists of factors that should be irrelevant for producing an effect (Meehl, 1967; Bird, 2018; Muthukrishna and Henrich, 2019). This would limit the need for replication due to the complexity and diversity of human subjects. But in almost all domains of psychology, it is difficult to construct or motivate such strong theory. With only weaker and little unified theory, most new experiments do not start with strong evidence about which moderators are

[9]Of course, there is a variety of quantitative and qualitative methods in psychological research, and qualitative methods are not always a good target for statistical analysis. But the question of whether the data are representative of the population of interest is important regardless of whether that data is quantitative or qualitative.

relevant or even about what they could be, and cannot make precise predictions for effect sizes. However, it is important to be precise about the form of such hypotheses: they may not make a precise prediction for an effect size, but they do postulate that this size is fixed (given the fixedness of relevant moderators). This makes both direct and conceptual replication crucial for supporting such hypotheses, for they allow more precise and triangulated tests of this hypothesized fixed value.[10]

Moving on to the availability and cost of evidence: Because psychological science concerns itself with human behavior and mental life, evidence is in one sense plentiful to collect but in another sense scarce. It is plentiful in the sense that there are scads of people around just about anywhere there's a psychology department. But it is scarce because of its locality and cost. The people available are most often psychology undergraduates; others—often also undergraduates—must be reimbursed for their time.[11] Unlike in disciplines dominated by big science, though, there is a more (but certainly not perfectly) equal distribution of access to new evidence among researchers, albeit at a relatively low rate, temporally, because of these aforementioned costs.[12] This means that lots of psychologists have access to a modest amount of data in a given time period relevant for the academic publication cycle. This in turn means that the results of their studies may be more susceptible to sampling error than studies with larger sample sizes and to lower generalizability than studies with more diverse sample sizes. These motivate direct replications testing each of these possibilities.

Finally, the incentive structure in academic psychology has many similarities with other academic disciplines, but also some differences (especially in social psychology) that contribute to the replication crisis. The reasons for this are complex and still somewhat contentious, but for present purposes only a few high-level observations are needed. Like other academic disciplines, psychology rewards novel discoveries, which overwhelmingly take the form of a statistically significant effect size (in a null hypothesis significance test) presented in a published study. By contrast, by and large experiments which do not find statistically significant effect sizes are not published. This leads to publication bias: the studies pub-

[10] Meehl (1967) wanted to distinguish this lack of precise predictions from the situation in physics, but perhaps overstated his case: there are many experimental situations in physics in which theory predicts the existence of an effect determined by an unknown parameter, too. Meehl (1967) was absolutely right, though, that one cannot rest simply with evidence against a non-zero effect size; doing so abdicates responsibility to find just what the aforementioned patterns of human behavior and mental life *are*.

[11] Online participant services such as Amazon Turk and other crowdsourced methods offer a potentially more diverse participant pool at a more modest cost (Uhlmann et al., 2019), but come with their own challenges.

[12] "Big science" is a historiographical cluster concept referring to science with one or more of the following characteristics: large budgets, large staff sizes, large or particularly expensive equipment, and complex and expansive laboratories (Galison and Hevly, 1992).

lished are not representative of the studies performed. In particular, the incidence of false positive findings is higher, meaning that published results are more likely to be affected by sampling error than their aggregate reported statistics would suggest if all studies were published. Moreover, because of the relatively low development and lack of unity of much psychological theory, relatively few of these discoveries build or depend upon each other, with the exception of work done within the same or allied labs.

These seem to be two reasons—but surely not the only, or most important—why there is (at least until recently) little incentive to identify or refrain from questionable research practices such as hypothesizing after the results are known ("HARKing"), excluding outlying data points based on personal judgment, not reporting all measured dependent variables or statistical tests performed, etc. This necessitates direct replications serving the last function thereof yet to be mentioned: testing that published results do not arise from questionable research practices or fraud. The procedure for this is similar for testing for sampling, except that it requires a different team for the replication than the one for the original study—i.e., it changes the "research team" contextual variable—and preferably uses a preregistered research protocol that describes the plan for data collection, evaluation, and analysis (Schmidt, 2017, p. 240).

4 A Top-Down Account: The Methodology of Research Programmes

It will be useful to contrast the bottom-up account from section 3 with a concrete example of a top-down account, one that begins with general principles about science or scientific virtues to deduce the importance of replication generally, hence psychology in particular. For this purpose, I will examine the account of Zwaan et al. (2018). As mentioned in the introductory section 1, they hold that "a finding needs to be repeatable to count as a scientific discovery" for Popperian reasons: "If a finding that was initially presented as support for a theory cannot be reliably reproduced using the comprehensive set of instructions for duplicating the original procedure, then the specific prediction that motivated the original research question has been falsified" (Zwaan et al., 2018, p. 2). For Popper, falsifiability is constitutive of scientific methodology, hence for Zwaan et al., replicability—in particular, direct replicability—is essential for a finding to count as a scientific discovery: "replicability is, in principle, an essential criterion for the effect to be accepted as part of the scientific literature ...and replication studies therefore evaluate the robustness of scientific findings" (Zwaan et al., 2018, p. 2). But they acknowledge the holism of testing: psychological hypotheses are

tested only in conjunction with a set of auxiliary hypotheses about the effects of contextual factors, the functioning of measuring instruments, etc. To accommodate this important and pervasive fact about scientific testing, they turn to MSRP.

Briefly,[13] in his MSRP, Lakatos (1970) considered a scientific research programme to consist of a sequence of theories with a shared "hard core," which is a collection of statements or hypotheses that themselves typically make no definite empirical prediction. The constancy of the hard core constitutes the identity condition for a research programme and reflects the commitments that scientists working in that programme are loathe to abandon. In order to produce empirical predictions, however, they must supplement the hard core with a "protective belt," a collection of auxiliary hypotheses that they are more willing to abandon—the so-called "negative heuristic." To the extent that a research programme, on balance, consists in a sequence of theories that predict novel and unexpected phenomena which are then corroborated by experiment, it is progressive and good science. To the extent that it fails these requirements, it is degenerating and bad science. Two common ways for this latter to happen are for the programme to fail to make hardly any definite novel predictions at all, and for its predictions to be falsified. This occurs when the conjunction of the hard core, the protective belt, and observations yield a contradiction.

The significance of this distinction for Zwaan et al. (2018) is that it allows them to identify what's wrong with research that always resorts to a new auxiliary hypothesis in the face of a failed replication, what Meehl (1990, p. 112) called "ad hockery":

> As more and more ad hockery piles up in the program, the psychological threshold (which will show individual differences from one scientist to another) for grave scepticism as to the hard core will be increasingly often passed, inducing an increasing number of able intellects to become suspicious about the hard core and to start thinking about a radically new theory.

Scientists who follow their skepticism in this way act rationally, while those who stick with degenerating research programs act irrationally. Replication, both direct and conceptual, thus provides a means to distinguish good psychological science from bad (Zwaan et al., 2018, p. 2).

There is something right in the identification by Zwaan et al. (2018) and Meehl (1990) of a problem with research that brushes off failed replications. But, as I argue in the remainder of this section, it can't quite be on the grounds of MSRP. I identify three problems:

1. MSRP fits poorly with the statistical hypotheses ubiquitous in psychological science.

[13]For secondary sources on MSRP, see Musgrave and Pigden (2016, §§2.2, 3.4)

2. As Feyerabend (1970, 1975, Ch. 16) first described, MSRP is normatively toothless, so can't ground any normative claims for replication in particular.

3. By following a top-down account, MSRP would make replication important in psychology by making it important in science generally, so that its importance follows from the same universal reasons. Yet there there are good local reasons to believe that replication is not as important in every scientific endeavor, or at least not for the same reasons.

Regarding the first problem, MRSP retains from Popper's falsificationism the essentially deductive character of scientific inference. The conjunction of the hard core and protective belt entail some novel empirical predictions; when those predictions are not borne out, modus tollens warrants inference to the negation of that conjunction. MSRP's "negative heuristic" then mandates inference to a negation of some hypothesis in the protective belt, but to no negations of hypotheses in the hard core. However, if the conjunction yields only (non-zero/one) probabilistic predictions, then no observation can conflict with them. Data may be quite (if not perfectly) unlikely, yet this warrants no inference via modus tollens. Logic thus demands no rejection of statistical auxiliary hypotheses. Research programmes employing them need never fear condemnation of degeneration.

Some (e.g., Gillies (1971)) have attempted to repair this defect by using significance levels as falsification thresholds for statistical hypotheses. According to this plan, a scientific community adopts a convention for how improbable or unexpected data can be until it is considered to furnish a falisfication. Advocates thereof might even point to the already deeply ingrained tradition of using a 0.05 significance level for this purpose. But the cure is worse than the poison, as it makes falsification—hence judgments about which research programmes are progressive and which are degenerating—conventional and liable to historical revision. This is not a remote possibility, as some have advocated lowering this threshold to 0.005 in response to the replication crisis (Benjamin et al., 2018), while others have insisted on determining the appropriate level on a case-by-case basis (Lakens et al., 2018). Such conventionality (or subjectivity, depending on the details of implementation) is inimical to the original goal of rationally grounding the importance of replication because it seems to undermine normative demands to replicate.

A similar conclusion arises from a traditional problem with MSRP.[14] It comes in the form of a dilemma (Feyerabend, 1970, 1975, Ch. 16). Either MSRP provides normative criteria for evaluating current science or it does not. If it does, then it condemns the history of science, for there are many historical examples of degenerating research programs that later

[14]For more on this, see Musgrave and Pigden (2016, §4).

became progressive. It seems that MSRP denies this horn: "In response to this, Lakatos distinguished appraisal from advice, and said that the task of the philosopher of science is to issue rules of appraisal, not to advise scientists (or grant-giving agencies) about what they ought to do. The Demarcation Criterion can evaluate the current state of play but it does not tell anyone what to do about it" (Musgrave and Pigden, 2016, §4). But if it does not provides normative criteria for evaluating current science, then it cannot condemn research programmes in psychological science for eschewing replication. Perhaps Zwaan et al. (2018) would therefore take the first horn of this dilemma instead, but in that case they would need to say more about how much ad hockery, exactly, is needed to make a research programme degenerate. Without this, the success or failure to replicate an important experiment cannot itself be indicative of a progressive or degenerate research programme, respectively, except perhaps only as a matter of personal taste.

The last criticism I consider may be related to a more general feature of MSRP, namely its ambition to structure scientific methodology regardless of field of study as a chapter in historicist rationality. However, there is significant evidence that there are descriptive and normative differences between scientific disciplines regarding methodology, including the importance of replication—direct replication, in particular (Chen, 1994; Norton, 2015; Leonelli, 2018). In his summary of some of this evidence for this, Guttinger (2020, p. 6) writes that whether replicability is necessary in a science may depend on "the type of questions researchers ask, the experimental setups they use, and the nature of the objects they analyse." Each of these provides a reason for why direct replication may not be essential in a scientific inquiry; in some cases, more than one of the reasons applies.[15]

Sometimes the research questions that scientists ask involve answers that are not intended to be replicable. Exploratory research provides an example of this. The results of strictly exploratory research are not typically confirmatory evidence for any particular scientific hypothesis or generalization. Rather, its results are indications about possibly fruitful research direction to pursue. The mark of good exploratory research, as difficult as it is to measure, is productive and valuable future results, of whatever sort it engenders, even if those results are not directly indicated by the original research. It is therefore less concerned with underdetermination, low-level or otherwise. Thus replicability is just the wrong concept to use for evaluating the results of exploratory research.

Sometimes the experimental or observational setup of a scientific inquiry makes direct replication prohibitive or impossible. This is the case in much of big science, where the scientific community often pools its resources to construct a single experimental apparatus.

[15]In what follows, I use my own examples rather than Guttinger's, with the exception of some overlap in discussion of Leonelli (2018).

The highest energy investigations at the Large Hadron Collider and thescope Human Genome Project are two notable examples, as are most examples of space and planetary physics that involve probes and satellites. In such cases a significant portion of a scientific community is involved in crafting the experimental apparatus, ensuring there will be enough data to eliminate practically the influence of sampling error, and in precluding questionable research practices. For example, in discussing the decades of preparations for the expected launch of Gravity Probe B (actually launched in 2004), Cartwright (1991, pp. 150–1) writes that direct replication

> is a guard against errors in our instruments. That is why it is not absolutely necessary. The more secure we are in the design of our instruments, the less need there is for [direct replication]. ...But the point is that it is worth doing. We expect that the design will be good enough to trust its outcomes, though we do not anticipate trying to reproduce them.

This is also the case for "research on materials that are rare, unique, perishable and/or inaccessible such as depletable samples stored in biobanks ...archaeological remains[,] or materials that are hard or expensive to access (such as very costly strains of transgenic mice)" (Leonelli, 2018, p. 137), even when the research questions involve trends and patterns of many events. As with the case of big science, "The onus of reproducibility shifts instead to the credibility and skills of the investigators entrusted with handling those materials" (Leonelli, 2018, p. 137). Much the same applies to risky medical research on people (and primate surrogate models).

Finally, sometimes the objects or phenomena in which scientists are interested themselves make replication prohibitive or impossible. This is the case with many historical, environmental, and social sciences. Archaeology, paleontology, climate science, cosmology, and parts of macroeconomics take as their respective objects of study the trends and causes within historically singular courses of events. Because of this, direct replication cannot be expected as a reasonable scientific demand. Medical, sociological, and anthropological researchers using ethnographic and other qualitative methods also do not expect direct replicability, not necessarily because of the singular nature of their study but because of the complexity of the textual data produced, the spatiotemporal localization of its origin, and the effects the researchers themselves have on their objects of study in the course of research (Leonelli, 2018, pp. 137–8).

Zwaan et al. (2018, p. 9) do anticipate these concerns, replying that "concerns about feasibility are orthogonal to the overarching value of direct replications for advancing scientific knowledge. The fact that replication studies are not always possible does not undermine

their value when they can be conducted." Some of the examples mentioned above show that this conclusion is not quite right. Replication studies with costs to human lives, heritage, or biodiversity may be possible, but their ethical costs may outweigh their scientific costs. The same can be seen just by focusing on scientific value. In big science, where the prolonged and widely distributed experimental design process and period of data collection is intended to fulfill the functions that direct replication can provide, direct replication would in fact be of little value. In any case, since Zwaan et al. (2018) concede that sometimes replications of valid scientific discoveries are not feasible, they must forfeit replication as a criterion for valid scientific discovery and for distinguishing progressive from degenerating research programmes.

5 Concluding Comparisons

The new, bottom-up account of the role of replication in psychological science that I provided in section 3 contrasts with the top-down account via MSRP in section 4 in three ways that I discuss presently. The first is on the the last point of criticism of MSRP: not just it, but top-down accounts in general seem to predict much more uniformity to replication's function and use across science than is observed. By contrast, bottom-up accounts predict that the functions that replication serves will vary from scientific discipline to discipline—indeed, even within scientific disciplines—correlated with the types of underdetermination challenges faced. Even superficial comparison with psychology of some of the examples mentioned, such as the big science of the Large Hadron Collider and the Human Genome Project or the some aspects of sciences interested in singular events such as archaeology and paleontology, reveals that the types of underdetermination faced—even low-level ones—vary across scientific investigations.

Second, and relatedly, top-down accounts require replication's universal import across science, and for the same reasons. By contrast, bottom-up accounts are *compatible* with replication performing some function or other in every science, but do not demand this as a matter of logic. Neither do they demand the converse. It seems to be conceptually possible that there could be scientific investigations that do not face any of the underdetermination challenges that are replications' function to assuage.[16]

Third, a top-down account, if correct, might be argued to posses a virtue of unity, in that it would show how the importance of replication follows from something like a universal

[16]Leonelli (2018) has argued that this possibility is realized in certain sciences that focus on qualitative data collection, but it is yet unclear whether this is really due to *pragmatic* limitations on the possibility of replications, rather than a lack of underdetermination, low-level or otherwise.

principle. In that sense it would provide an explanation of replication's importance and role by unification. Bottom-up accounts do not have the same sort of unity, but the account I have advanced, based on a functional characterization of replication, possesses unity nonetheless. It understand replication as a response to underdetermination at the level of individual or small groups of scientific studies, even though the exact nature of the underdetermination challenge varies from case to case depending on the epistemic, material, and institutional features of the scientific discipline at hand. It unifies in the same sense that functional accounts of other phenomena unify: not under the aegis of a general principle, but according to the type of role played across the phenomena in question. Replications indeed serve a unified function, even if realized in quite different ways.

References

Bailer-Jones, D. M. (2009). *Scientific Models in Philosophy of Science*. University of Pittsburgh Press, Pittsburgh, PA.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604):452–454.

Begley, C. G. and Ellis, L. M. (2012). Raise standards for preclinical cancer research: Drug development. *Nature*, 483(7391):531–533.

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, 100(3):407.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., et al. (2018). Redefine statistical significance. *Nature Human Behaviour*, 2(1):6.

Bird, A. (2018). Understanding the replication crisis as a base rate fallacy. *The British Journal for the Philosophy of Science*, forthcoming.

Bogen, J. and Woodward, J. (1988). Saving the phenomena. *The Philosophical Review*, 97(3):303–352.

Brewer, W. F. and Chinn, C. A. (1994). Scientists' responses to anomalous data: Evidence from psychology, history, and philosophy of science. In *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, volume 1, pages 304–313. Philosophy of Science Association.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., and Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14(5):365–376.

Callaway, E. (2011). Report finds massive fraud at Dutch universities. *Nature*, 479(7371):15.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., et al. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280):1433–1436.

Carpenter, S. (2012). Government sanctions Harvard psychologist. *Science*, 337(6100):1283–1283.

Cartwright, N. (1991). Replicability, reproducibility, and robustness: comments on Harry Collins. *History of Political Economy*, 23(1):143–155.

Chen, X. (1994). The rule of reproducibility and its applications in experiment appraisal. *Synthese*, 99:87–109.

Dunlap, K. (1925). The experimental methods of psychology. *The Pedagogical Seminary and Journal of Genetic Psychology*, 32(3):502–522.

Dunlap, K. (1926). The experimental methods of psychology. In Murchison, C., editor, *Psychologies of 1925: Powell Lectures in Psychological Theory*, pages 331–351. Clark University Press, Worcester, MA.

Feest, U. (2019). Why replication is overrated. *Philosophy of Science*, 86(5):895–905.

Feyerabend, P. (1970). Consolation for the specialist. In Lakatos, I. and Musgrave, A., editors, *Criticism and the Growth of Knowledge*, pages 197–230. Cambridge University Press., Cambridge.

Feyerabend, P. (1975). *Against Method*. New Left Books, London.

Fidler, F. and Wilcox, J. (2018). Reproducibility of scientific results. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2018 edition.

Franklin, A. and Howson, C. (1984). Why do scientists prefer to vary their experiments? *Studies in History and Philosophy of Science Part A*, 15(1):51–62.

Galison, P. and Hevly, B. W., editors (1992). *Big Science : The Growth of Large-Scale Research*. Stanford University Press, Stanford, CA.

Gelman, A. (2018). Don't characterize replications as successes or failures. *Behavioral and Brain Sciences*, 41:e128.

Gillies, D. A. (1971). A falsifying rule for probability statements. *The British Journal for the Philosophy of Science*, 22(3):231–261.

Gómez, O. S., Juristo, N., and Vegas, S. (2010). Replications types in experimental disciplines. In *Proceedings of the 2010 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement*, ESEM '10, New York, NY, USA. Association for Computing Machinery.

Greenwald, A. G., Pratkanis, A. R., Leippe, M. R., and Baumgardner, M. H. (1986). Under what conditions does theory obstruct research progress? *Psychological Review*, 93(2):216–229.

Guttinger, S. (2020). The limits of replicability. *European Journal for Philosophy of Science*, 10(10):1–17.

Hagger, M. S., Chatzisarantis, N. L., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., Brand, R., Brandt, M. J., Brewer, G., Bruyneel, S., et al. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science*, 11(4):546–573.

Hansson, S. O. (2013). Defining pseudoscience and science. In Pigliucci, M. and Boudry, M., editors, *Philosophy of Pseudoscience: Reconsidering the Demarcation Problem*, pages 61–77. University of Chicago Press, Chicago.

Harris, T. (2003). Data models and the acquisition and manipulation of data. *Philosophy of Science*, 70(5):1508–1517.

Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In Lakatos, I. and Musgrave, A., editors, *Criticism and the growth of knowledge*, pages 91–196. Cambridge University Press, Cambridge.

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A., Argamon, S. E., Baguley, T., Becker, R. B., Benning, S. D., Bradford, D. E., et al. (2018). Justify your alpha. *Nature Human Behaviour*, 2(3):168.

Laudan, L. (1983). The demise of the demarcation problem. In Cohan, R. and Laudan, L., editors, *Physics, Philosophy, and Psychoanalysis*, pages 111–127. Reidel, Dordrecht.

Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., Carter, S. L., Stewart, C., Mermel, C. H., Roberts, S. A., et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457):214–218.

Laymon, R. (1982). Scientific realism and the hierarchical counterfactual path from data to theory. In *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, volume 1, pages 107–121. Philosophy of Science Association.

LeBel, E. P., Berger, D., Campbell, L., and Loving, T. J. (2017). Falsifiability is not optional. *Journal of Personality and Social Psychology*, 113(2):254–261.

Leonelli, S. (2018). Rethinking reproducibility as a criterion for research quality. In Boumans, M. and Chao, H.-K., editors, *Including a Symposium on Mary Morgan: Curiosity, Imagination, and Surprise*, volume 36B of *Research in the History of Economic Thought and Methodology*, pages 129–146. Emerald Publishing Ltd.

Leonelli, S. (2019). What distinguishes data from models? *European Journal for Philosophy of Science*, 9(2):22.

Machery, E. (2020). What is a replication? *Philosophy of Science*, forthcoming.

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, 34(2):103–115.

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry*, 1(2):108–141.

Musgrave, A. and Pigden, C. (2016). Imre Lakatos. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2016 edition.

Muthukrishna, M. and Henrich, J. (2019). A problem in theory. *Nature Human Behaviour*, 3(3):221–229.

Norton, J. D. (2015). Replicability of experiment. *THEORIA. Revista de Teoría, Historia y Fundamentos de la Ciencia*, 30(2):229–248.

Nosek, B. A. and Errington, T. M. (2017). Reproducibility in cancer biology: Making sense of replications. *Elife*, 6:e23383.

Nosek, B. A. and Errington, T. M. (2020). What is replication? *PLoS biology*, 18(3):e3000691.

Nuijten, M. B., Bakker, M., Maassen, E., and Wicherts, J. M. (2018). Verify original results through reanalysis before replicating. *Behavioral and Brain Sciences*, 41:e143.

Open Science Collaboration (OSC) (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716.

Popper, K. R. (1959). *The Logic of Scientific Discovery*. Routledge, Oxford.

Radder, H. (1992). Experimental reproducibility and the experimenters' regress. In *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, volume 1, pages 63–73. Philosophy of Science Association.

Rosenthal, R. (1990). Replication in behavioral research. In Neuliep, J. W., editor, *Handbook of Replication Research in the Behavioral and Social Sciences*, volume 5 of *Journal of Social Behavior and Personality*, pages 1–30. Select Press, Corte Madera, CA.

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology*, 13(2):90–100.

Schmidt, S. (2017). Replication. In Makel, M. C. and Plucker, J. A., editors, *Toward a More Perfect Psychology: Improving Trust, Accuracy, and Transparency in Research*, pages 233–253. American Psychological Association.

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science*, 9(1):76–80.

Simons, D. J., Shoda, Y., and Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science*, 12(6):1123–1128.

Stanford, K. (2017). Underdetermination of Scientific Theory. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2017 edition.

Suppes, P. (1962). Models of data. In Nagel, E., Suppes, P., and Tarski, A., editors, *Logic, Methodology and Philosophy of Science: Proceedings of the 1960 International Congress*, pages 252–261. Stanford University Press, Stanford, CA.

Suppes, P. (2007). Statistical concepts in philosophy of science. *Synthese*, 154(3):485–496.

Uhlmann, E. L., Ebersole, C. R., Chartier, C. R., Errington, T. M., Kidwell, M. C., Lai, C. K., McCarthy, R. J., Riegelman, A., Silberzahn, R., and Nosek, B. A. (2019). Scientific utopia III: Crowdsourcing science. *Perspectives on Psychological Science*, 14(5):711–733.

Zwaan, R. A., Etz, A., Lucas, R. E., and Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences*, 41:e120.